SOURCE DETECTION AND ANALYSIS

X-Ray Astronomy School, Wallops Island, 2003 May 12-15

Vinay Kashyap

Why source detection?

Tieing up loose ends:
Source detection in practice

Opening new can of worms:
Dealing with detected sources

- Hardness Ratios
- $\log N - \log S$

Aneta Siemiginowska and Peter Freeman spoke about Image analysis and the theory of source detection. Here we borrow heavily from the concepts and terminology introduced by them to describe, first, source detection in practice and second, what to do with these sources once detected.

Source Detection in the Real World: Snares and Traps

1. The Starter Kit
   - Aspect dither
   - Bad pixels
   - Exposure maps
2. Here be Hippogrif
   - Point Spread Functions
   - Varying background
   - Extended Sources
   - Overlapping Sources
3. What, me worry?
   - Source Position and plate scale
   - Source spectrum and ExpMaps and PSFs
   - Pileup
   - Detection Sensitivity (Type I Errors)
   - Detection Probability (Type II Errors)

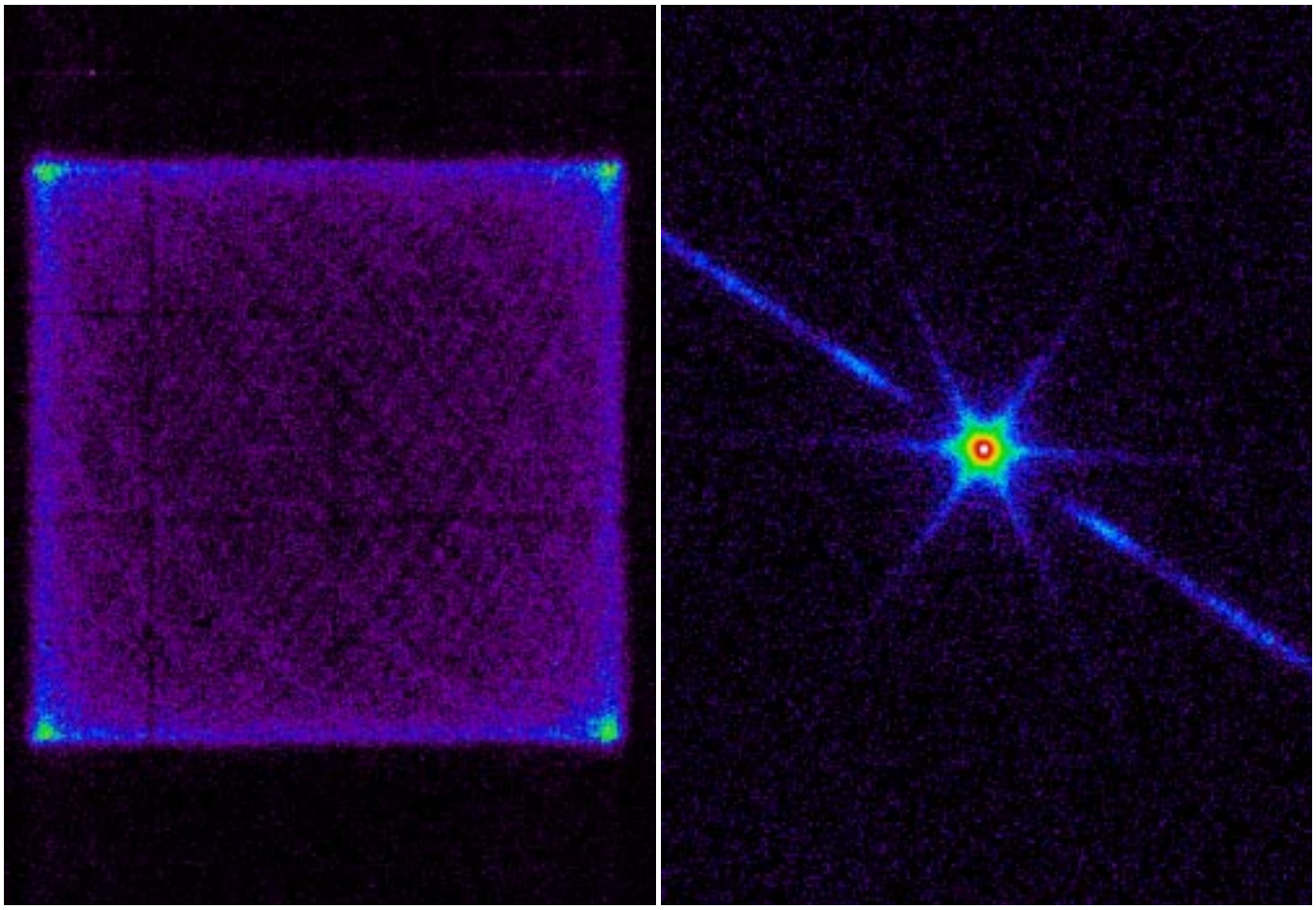

Figure 1: Capella HRC-S/LETG 0^{th} order image. *Left:* pre dither correction, in detector coordinates. *Right:* post dither correction, in sky coordinates.

Unlike optical telescopes which essentially take photographs of the sky, X-ray telescopes record photon events and are not obliged to keep pointing exactly towards the source. In fact, all X-ray telescopes *dither*. The counts recorded in detector coordinates are smeared out and must be reconstructed from the (deterministic, i.e., not statistical) aspect solution into sky coordinates. Source detection must be performed in sky coordinates.

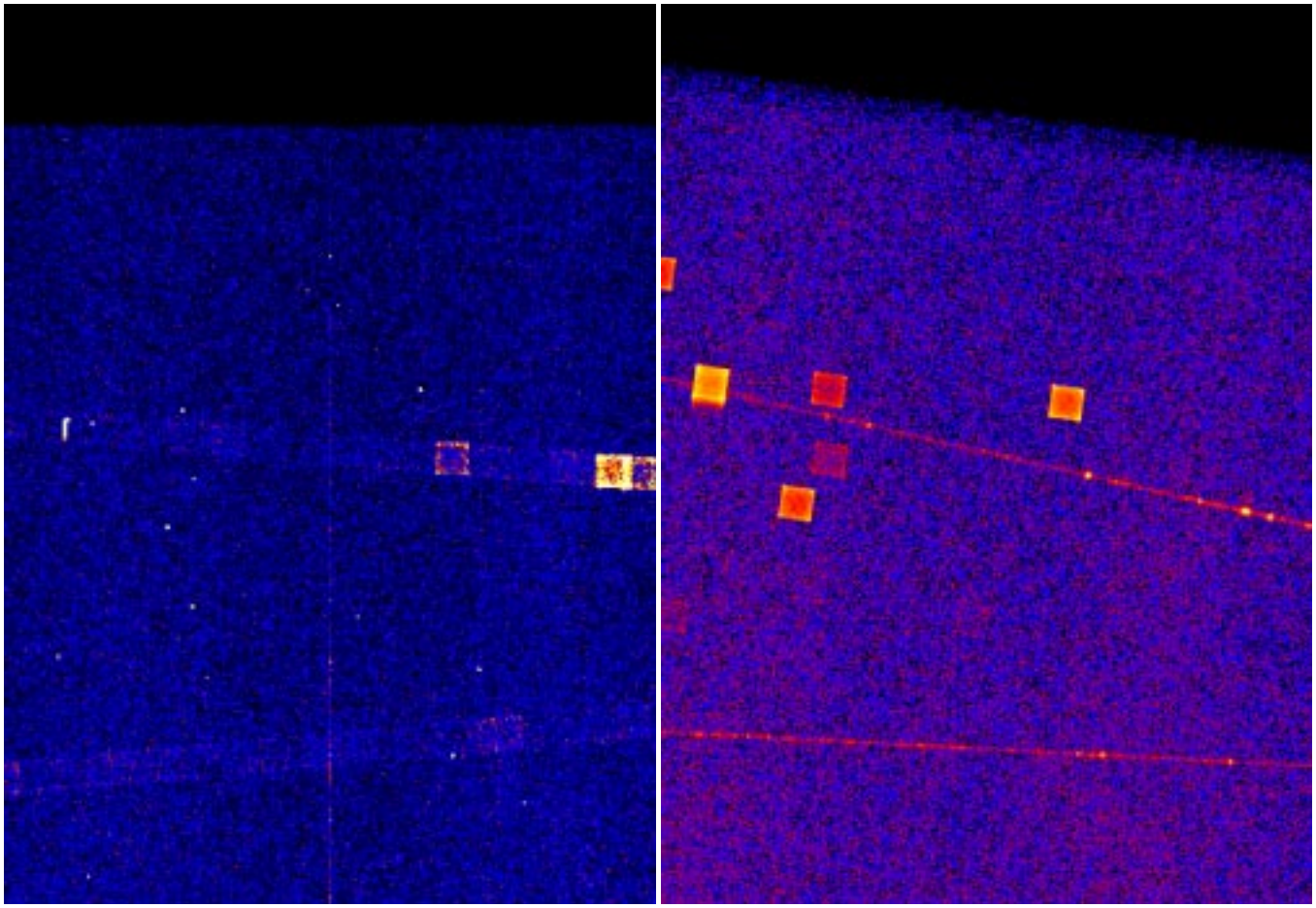

Figure 2: Capella ACIS-S/HETG image of dispersed HEG and MEG spectra. *Left:* pre dither correction, showing dither pattern at location of strong lines and point-like hot pixels. *Right:* post dither correction, showing reconstructed dispersed gratings spectra and "undithered" hot pixels.

Hot pixels on the detector produce a characteristic image in sky coordinates which will contaminated source detection. These bad pixels must be removed from the data prior to running any detection algorithm.

Removal of hot pixels leaves holes in the image. In order that detection algorithms not flag them as "negative" sources, they must be told that the exposure at these pixels is zero – this is done via exposure maps. Exposure maps are images containing information on how long a given pixel was exposed to the sky, and include factors such as the field of view, observation time, dead time, vignetting, etc. Normally, exposure maps are given in units of [s], though *Chandra* maps are in units of [cm^2]. If not supplied, detection algorithms are unable to adjust to the edge of the FOV and tend to report a large number of astrophysically spurious sources.

How do source detection algorithms include the information in an exposure map? It is usually a bad idea to simply divide the data image by the exposure map. Instead, `wavdetect` handles it by computing the response of the exposure map to the mexican-hat wavelet and removing this from the observed reponse.

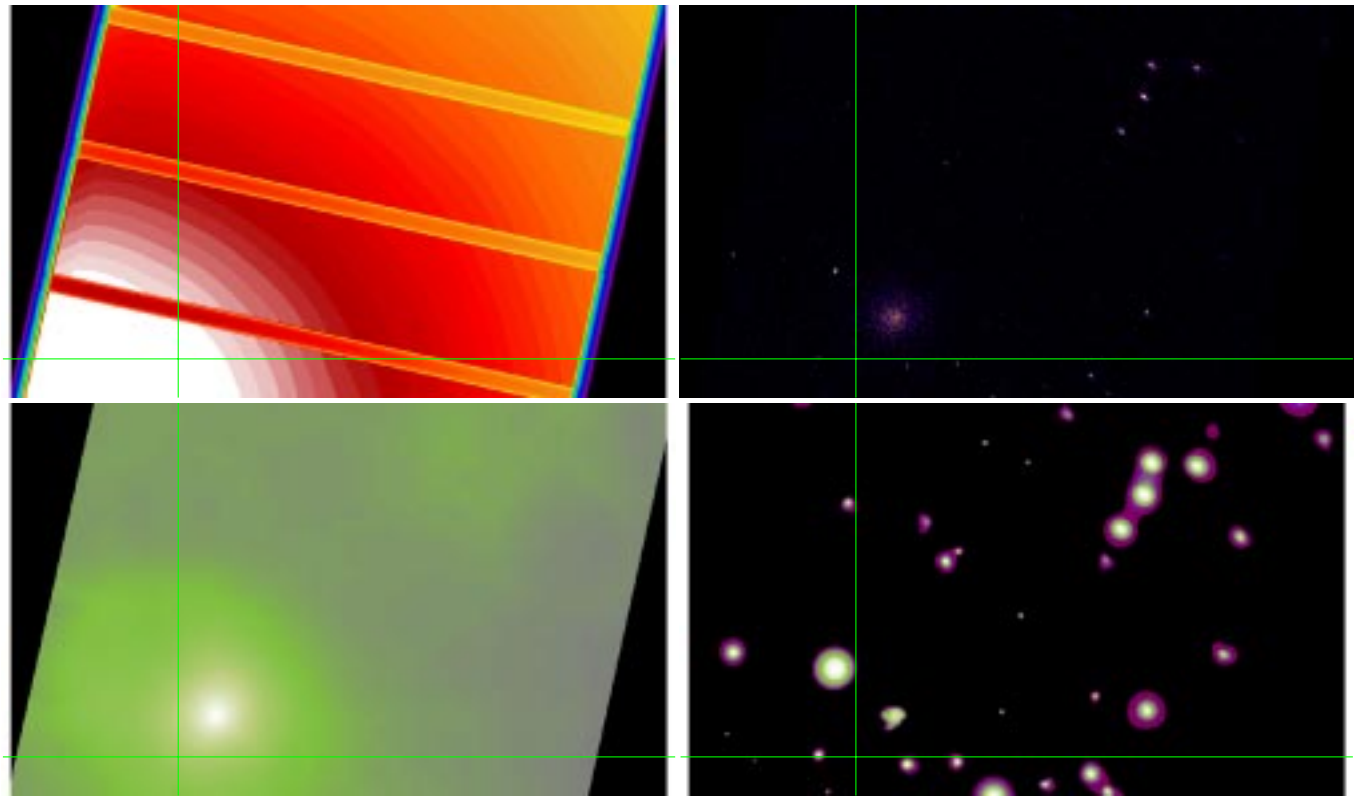

Figure 3: MS 1137 ACIS-I image. *Top Left:* Exposure map (note vignetting and removed bad columns). *Top Right:* Data image, with extended source (the cluster MS 1137) in top right quadrant of cross-hairs. *Bottom Left:* Background computed for source detection via `wavdetect` (note "bump" due to extended source, but the dynamic range in this image is <2x). *Bottom Right:* Sources detected by `wavdetect`. Dynamic range is $> 10^3$.

It is an intractable problem to estimate the background directly underneath a source, especially an extended source. We can only expect to minimize it.

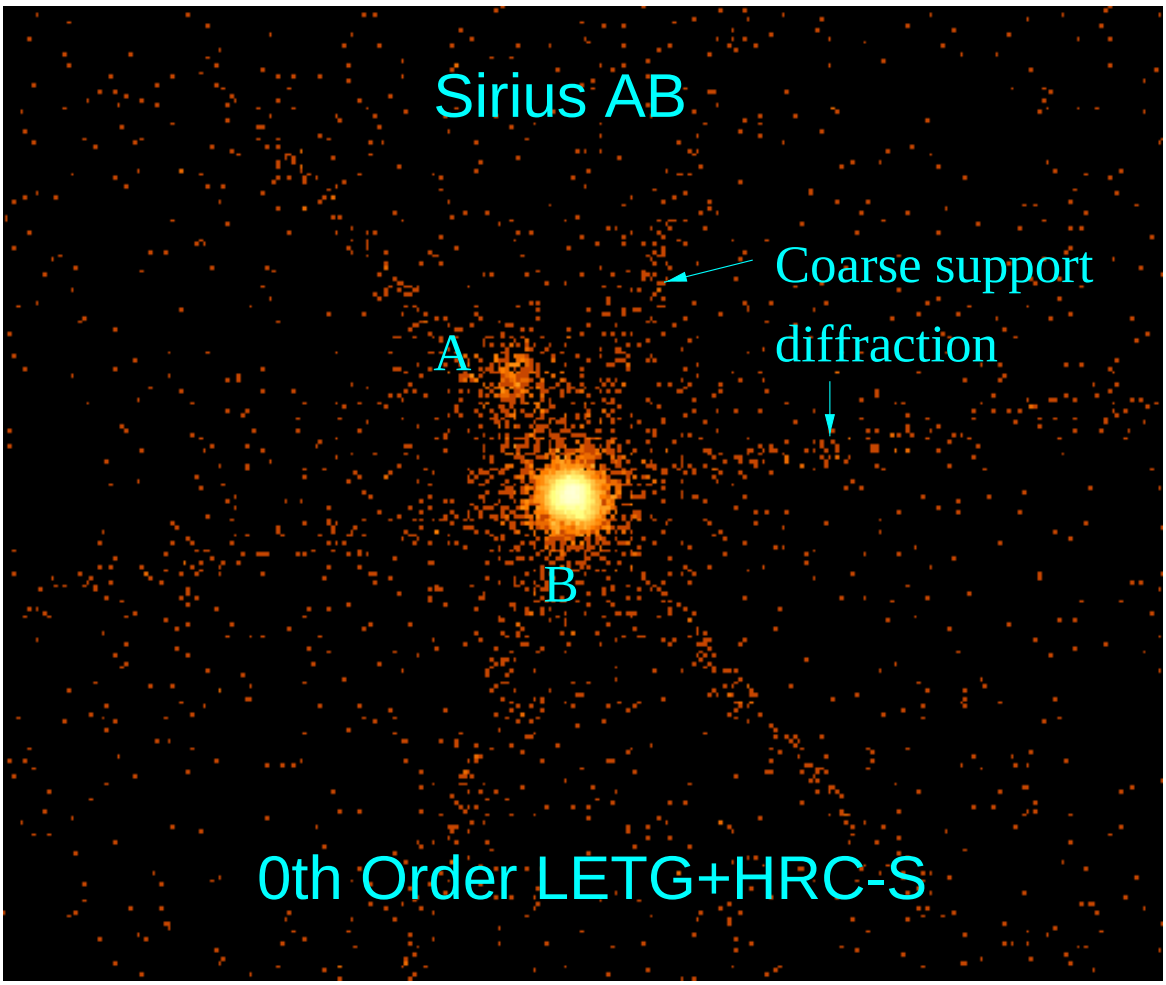

Figure 4: Sirius A and B, HRC-S/LETG 0^{th} order image. Sirius A is not an X-ray emitter, and is visible only because of a UV leak.

While detection algorithms such as `wavdetect` are designed to handle the large dynamic range in X-ray data and to search for faint sources close to strong sources, and sources of comparable strength that may be close to each other, the algorithms are not perfect, and visual inspection is necessary to ensure no sources have been missed.

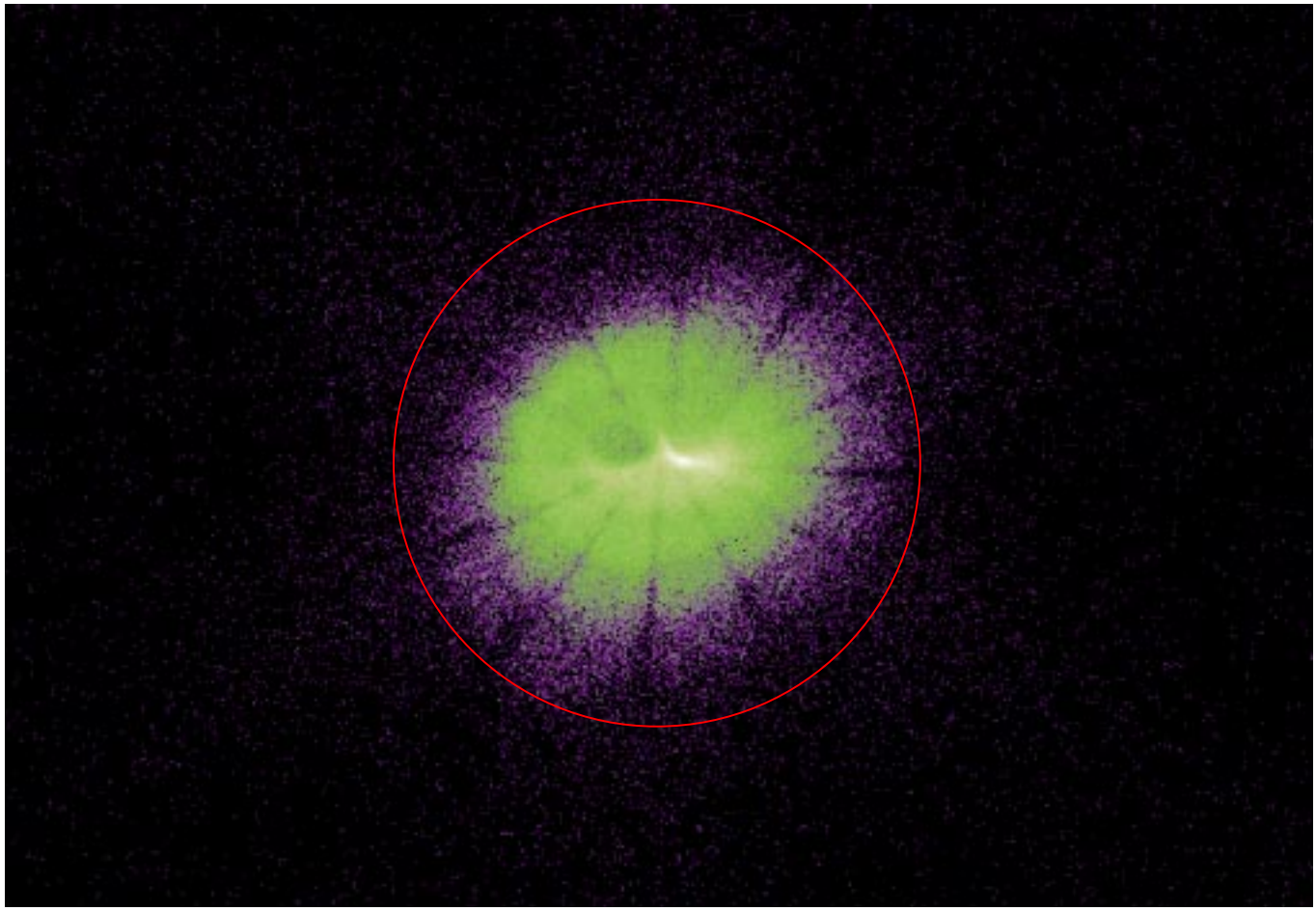

Figure 5: HZ 43 at 10 arcmin off-axis, with HRC-S/LETG.

Because point spread functions (PSFs) are highly non-Gaussian in generally non-smooth, especially at large off-axis angles, it is always a good idea to visually double check the detected source list against the image. Sometimes a single source with a complicated PSF may get detected as a "double" source.

Often, a simple source centroiding algorithm that may be used to determine the location of the source may be too heavily influenced by the surrounding background region to move the centroid to the wrong place (there is a fix in the latest version of `wavdetect` to mitigate this effect).

It is also a good idea to double check the locations of all detected sources against other cataloged coordinates wherever available, for all the sources in the field, regardless of whether they are your primary target or not. Spacecraft aspect corrections are generally very good, but nevertheless, it is very useful to do this check, especially in crowded fields.

When computing an exposure map, make sure you also include chip-to-chip variation via some spectral dependence, otherwise the boundary between e.g., the BI and FI chips of ACIS-S will get flagged as a source.

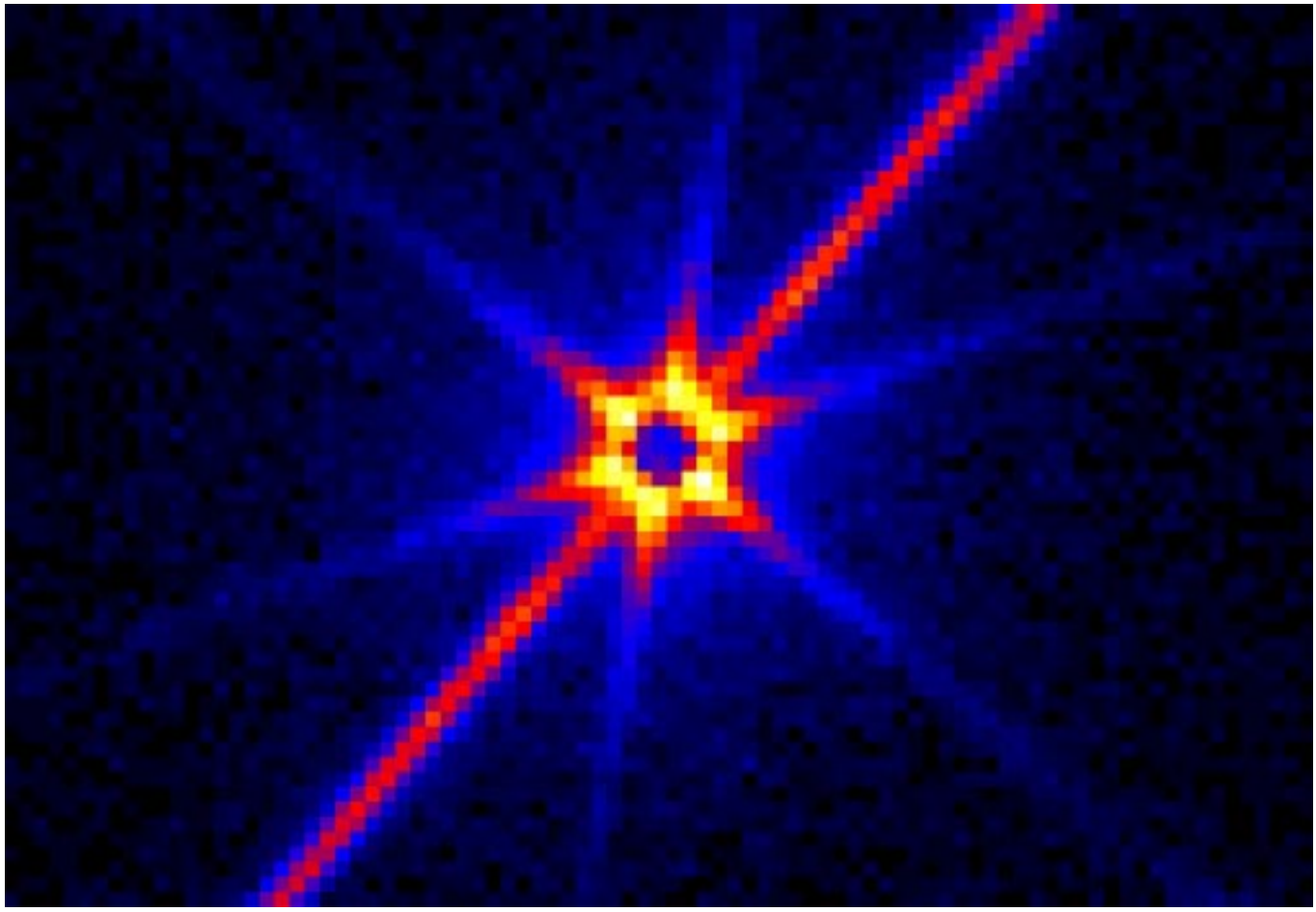

Figure 6: XTE J1118 ACIS-S/LETG 0^{th} order image, showing the dramatic effect of pileup on the central portion of the PSF.

Peter Freeman dealt with the issue of the detection threshold as the probability of rejecting the null hypothesis that the observed counts may be explained as a fluctuation from the background.

Note that this does not impute a significance value for *detected* sources.

It used to be that to detect a source, its flux would be measured and the source would be declared detected if the measured flux were statistically different (e.g., at 3σ) from 0. In `wavdetect` and similar new algorithms, detection is divorced from flux measurement. Hence, the column SRC_SIGNIFICANCE in the output source list from `wavdetect`, which computes S/N of the detected sources post hoc, is useless and is not a part of the detection process. No filtering of the source list is to be performed using these values, because if that is done then the computed detection probabilities will not apply, and the adopted detection threshold becomes meaningless.

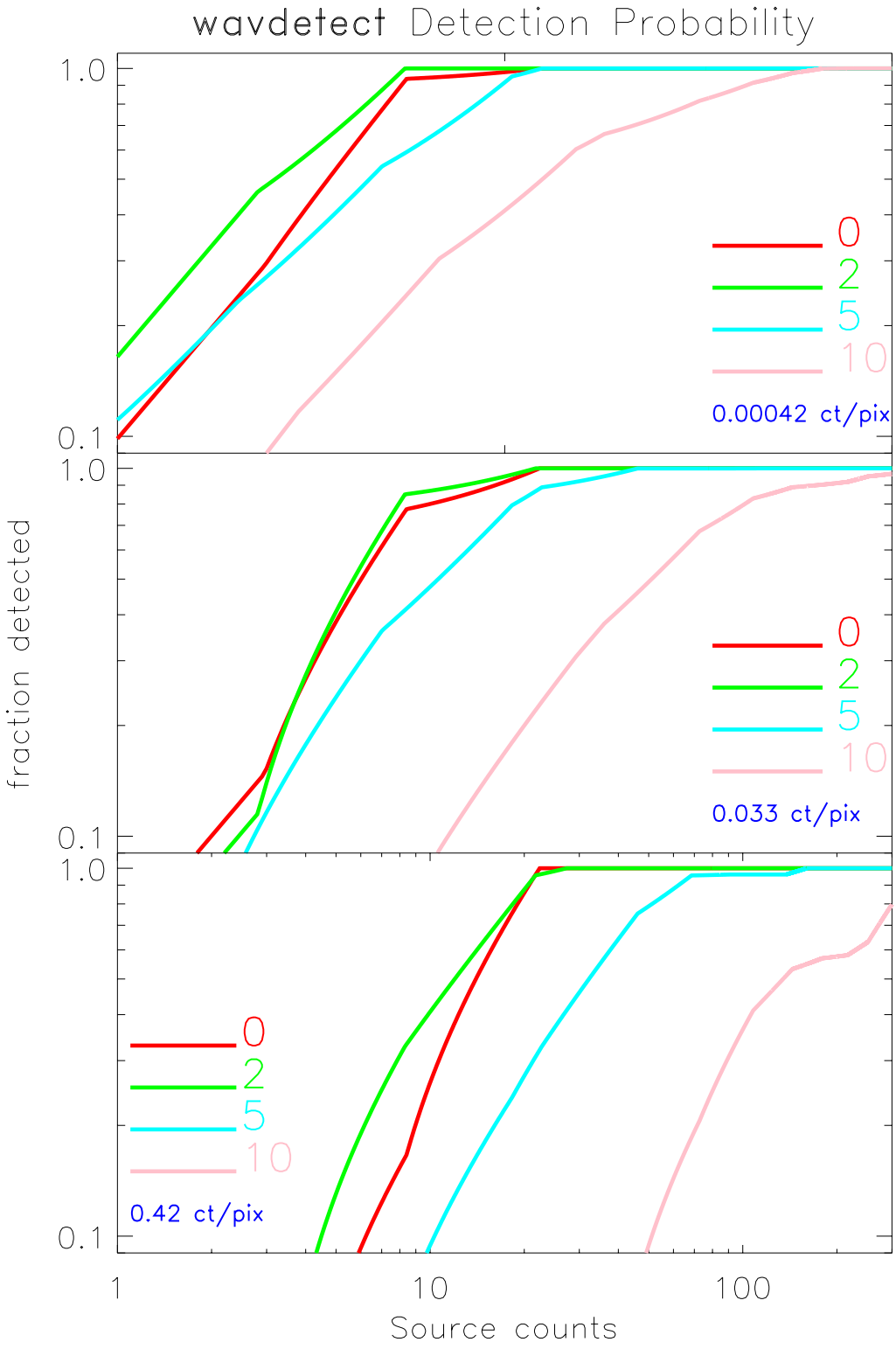

Figure 7: Probability of detection of a source with **wavdetect** as a function of intrinsic source strength, computed via simulations. The top plot is for a low background, the middle plot is for a medium background, and the bottom plot is for a high background. Within each plot, the different lines are for sources located at different off-axis angles (on-axis, 2′, 5′, and 10′ off-axis).

The Type I and Type II erros **must** be taken into account while computing, e.g., the $\log(N > S) - \log(S)$.

Hardness Ratios
(Not as simple as they seem)

Useful with large samples and when spectral fits are unfeasible; can separate out different source populations

Types of hardness ratios:

- $R = \frac{S}{H}$,
  $\sigma_R = R\sqrt{\left(\frac{\sigma_S}{S}\right)^2 + \left(\frac{\sigma_H}{H}\right)^2}$

- $C = \log_a(S) - \log_a(H)$,
  $\sigma_C = \frac{1}{\log(a)}\sqrt{\left(\frac{\sigma_S}{S}\right)^2 + \left(\frac{\sigma_H}{H}\right)^2}$

- $HR = \frac{H-S}{H+S}$,
  $\sigma_{HR} = \frac{2}{(H+S)^2}\sqrt{H^2\sigma_S^2 + S^2\sigma_H^2}$

Caveats

- Mathematical obstinacy
- Poisson statistics
- Background
- Upper limits

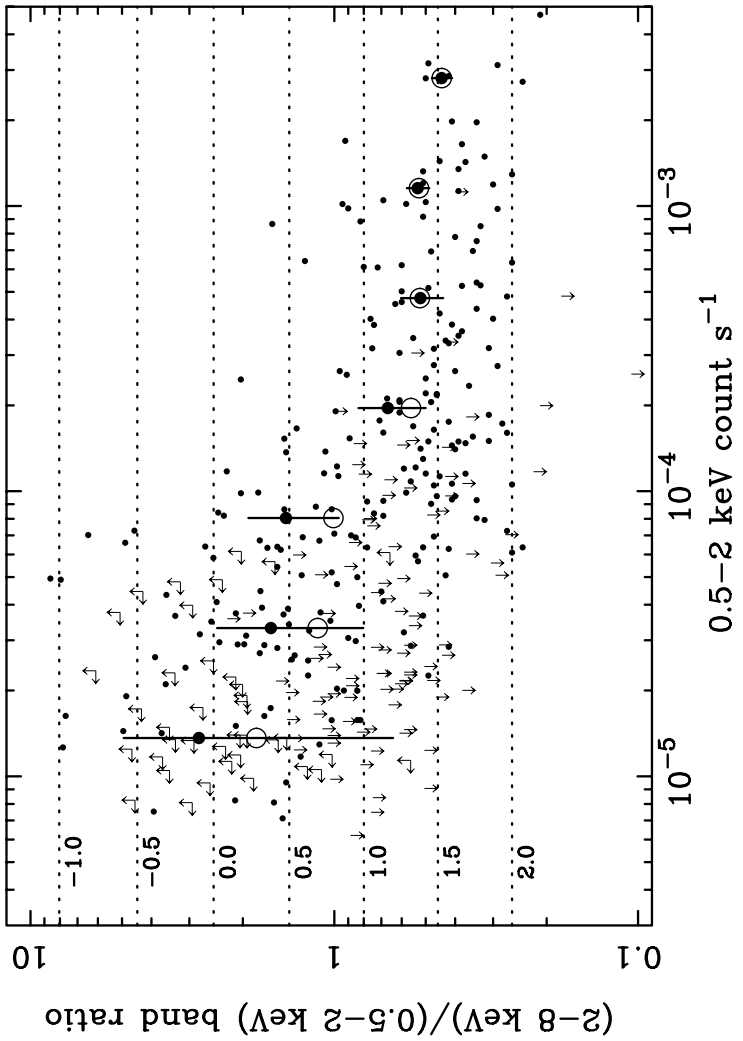

Fig. 12.— Band ratio as a function of soft-band count rate for the sources in Table 3. Small solid dots show sources detected in both the soft and hard bands, and plain arrows show sources detected in only one of these two bands (sources detected in only the full band are not plotted). To reduce symbol crowding, we do not show error bars for each of the small solid dots. Instead, the large solid dots show average band ratios and error bars for the small solid dots as a function of soft-band count rate (these large solid dots are given only to show the size of the errors, and they should not be interpreted statistically). The open circles show average band ratios derived from stacking analyses (following §3.3 of Paper VI). Horizontal dotted lines are labeled with the photon indices that correspond to a given band ratio assuming only Galactic absorption (these were determined using the CXC's Portable, Interactive, Multi-Mission Simulator).

Hardness ratios can be used to infer the existence of distinct source populations, as e.g., in the Chandra Deep Field North data, where Brandt et al. find that the hardness increases for faint sources. But is this effect real? The probability distributions of hardness ratios in the low-counts limit have some very unintuitive behaviors.

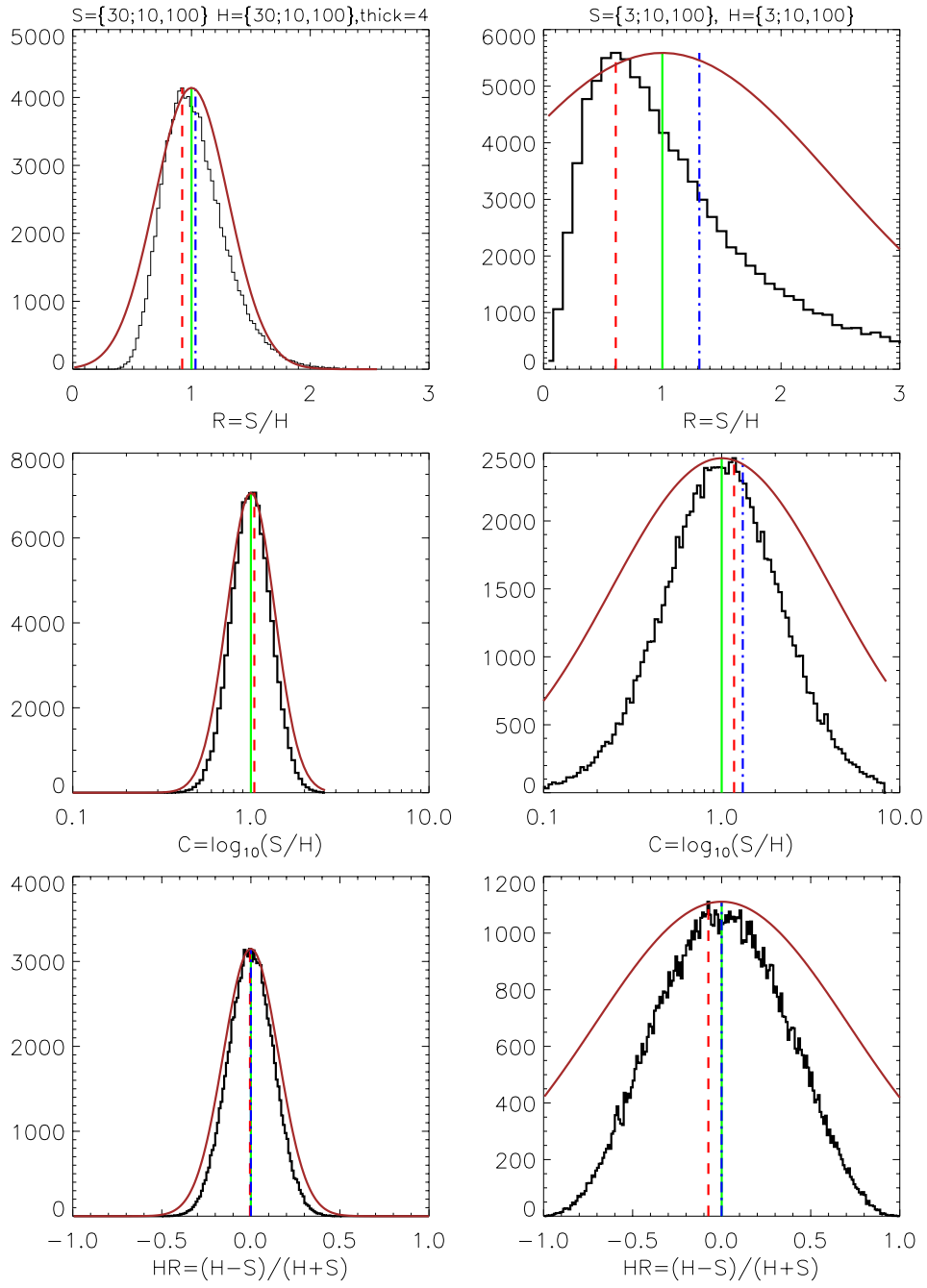

Figure 8: Probability distributions of different types of hardness ratios for high counts (*left column*; 30 source counts, and 10 background counts in an area 100 times larger than the source area) and low counts (*right column*; 3 source counts, and the same background as before). Top 2 panels are for the simple ratio $R = S/H$, the middle 2 panels are for the color, $C = \log_{10}(S/H)$, and the bottom 2 panels are for the fractional difference, $HR = (H-S)/(H+S)$. The black histogram is the distribution derived from a Monte Carlo simulation. The mode is marked by the red dashed line, the classical expected value by the solid green line, and the mean by the blue dot-dashed line. The brown curves are Gaussians centered on the classical values with σ computed by error propagation.

The Gaussian approximation is generally adequate in the high counts limit, but fails spectacularly for low counts. The simple ratio is heavily skewed, with the most probable value being much lower than the expected value and the mean being usually undefinable. The color and fractional differences are better behaved, though in the low-counts limit the errors are not properly propagated. Software to do these calculations will soon be available within CIAO.

$\log N - \log S$

Say $n(r) = n_0$, $f(L_x) = \delta(L_x - L_{x0})$, $S = \frac{L_{x0}}{4\pi r^2}$.
Number of sources within r,
$N(< r) = \frac{4\pi}{3} r^3 n_0$, or $N(> S) = n_0 \frac{4\pi}{3} \left(\frac{L_{x0}}{4\pi S}\right)^{\frac{3}{2}}$, i.e.,
$N(> S) \propto S^{-\frac{3}{2}}$.

In general,

$$N(> S, l, b) = d\Omega \int_0^\infty dL_x \int_0^\infty dL'_x f(L'_x) \sigma(L_x, L'_x) \int_0^{r'} n(r, l, b) r^2 dr$$

where r' is implicitly defined by $S = \frac{L'_x}{4\pi r'^2} e^{-\tau(r')}$.

Biases:

- Source confusion
- False sources (false positives)
- Lost sources (false negatives)
- Malmquist Bias
- Faint source fluctuations
- Eddington Bias

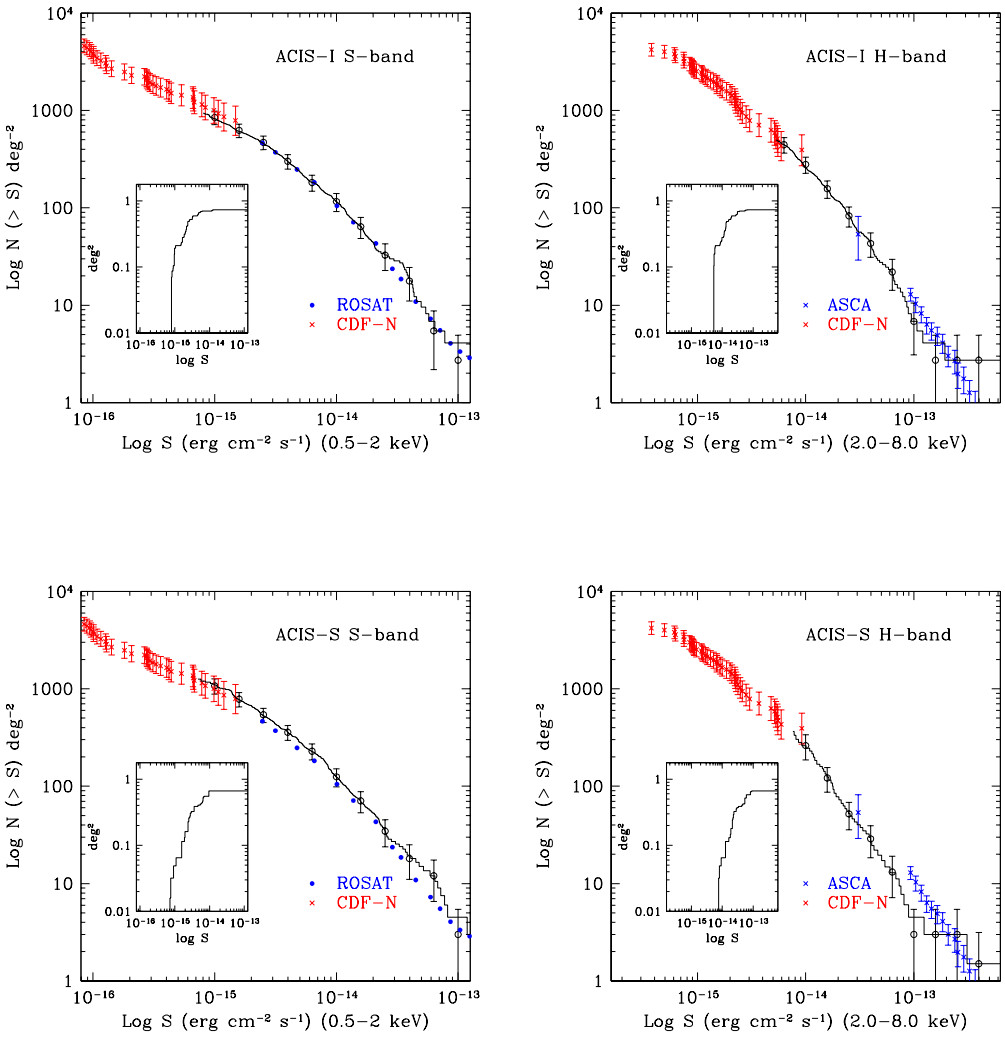

Figure 14. Log(N)–Log(S) in the soft energy band (0.5-2 keV; left panel) and in the hard band (2.0-8.0 keV; right panel), measured separately in ACIS-I (ccdid=0-3; top panel) and ACIS-S (ccdid=7; bottom panel). Also plotted are results from Hubble Deep Field-North and previous missions (ROSAT in the S band and ASCA data in the H band).

Biases:

- Source confusion
  ⋆ in crowded fields, detection algorithms tend to miss weaker sources near strong sources, and sometimes they merge multiple sources into a single one
- False sources (false positives)
  ⋆ false sources that arise due to fluctuations in the background will contaminate the numbers at the low flux end
- Lost sources (false negatives)
  ⋆ because of statistical fluctuations, sources of a given intrinsic strength will produce different numbers of counts at any given observation, and as the source strength becomes smaller, the chances of it not being detected increase even if nominally it is above the detection threshold
- Malmquist Bias
  ⋆ the volume in which high-luminosity sources can be detected is larger than the volume in which low-luminosity sources are detected; thus luminous objects will be overrepresented in flux-limited samples
- Faint source fluctuations
  ⋆ when there are larger numbers of low-flux sources than high-flux sources, then statistical fluctuations result in a larger number of the weaker sources deflected into higher flux regimes than stronger sources that are deflected into lower flux regimes
- Eddington Bias
  ⋆ when fluxes of sources with intensities near the detection threshold are measured, there is a tendency for the average measured flux to be higher than the true fluxes, because of the fact that fluctuations towards smaller counts will be censored out because of the detection threshold